Quality & Quantity
https://doi.org/10.1007/s11135-020-00992-w

Dynamics and tipping point of issue attention in newspapers: quantitative and qualitative content analysis at sentence level in a longitudinal study using supervised machine learning and big data

A. E. Opperhuizen[1,2] · **K. Schouten**[1,3]

Abstract

This study aims to provide a more sensitive understanding of the dynamics and tipping points of issue attention in news media by combining the strengths of quantitative and qualitative research. The topic of this 25-year longitudinal study is the volume and the content of newspaper articles about the emerging risk of gas drilling in The Netherlands. We applied supervised machine learning (SML) because this allowed us to study changes in the quantitative use of subtopics at the detailed sentence level in a large number of articles. The study shows that the actual risk of drilling-induced seismicity gradually increased and that the volume of newspaper attention for the issue also gradually increased for two decades. The sub-topics extracted from media articles during the low media attention period, covering factual information, can be interpreted as a part of episodic frame patterns about the drilling and its consequences. However, a sudden major shift in newspaper attention can be observed in 2013. This sudden disjointed expansion in the volume of media attention on this large-scale technology occurred after a governmental authority classified the drilling-induced earthquakes as a safety issue. After the disjointed issue expansion, *safety* and *decision making* were the main subtopics linked to the thematic frames, *responsibility*, *conflict*, *human interest*, and *morality*. We conclude that SML is a promising tool for future analysis of the growing number of publicly available digitalized textual big datasets, particularly for longitudinal studies and analysis of tipping points and reframing.

Keywords Media attention · Supervised machine learning · Risk · Tipping point · Reframing

✉ A. E. Opperhuizen
opperhuizen@essb.eur.nl

K. Schouten
schouten@ese.eur.nl

[1] Erasmus School of Social and Behavioural Sciences, Rotterdam, The Netherlands

[2] Social and Behavioural Sciences, Erasmus University Rotterdam, Oudlaan 50, Room T17-54, P.O. Box 1738, 3062 PA Rotterdam, The Netherlands

[3] Erasmus School of Economics, Erasmus University Rotterdam, Oudlaan 50, Room W-H8-09, P.O. Box 1738, 3062 PA Rotterdam, The Netherlands

Published online: 20 May 2020

1 Introduction

Media serve as the gatekeepers of information for the general public and fulfil an essential role by informing citizens about the risks and benefits of activities or situations (Shoemaker and Schäfer 1996; Schäfer 2012). Gruszczynski and Wagner (2017) argued, after an analysis of more than 400 media studies, that media coverage of a topic predicts citizens' attention on that same issue and raises awareness. Consequently, limited, or a lack of, media coverage may contribute to unawareness. For example, Kahlor et al. (2019) reported that citizens of Texas are mostly unaware of induced seismicity related to the extraction of gas and oil. Furthermore, Fisk et al. (2017) reported that the media coverage of this induced seismicity was limited for many years and that media frames emphasized the economic importance of oil and gas production, with little attention paid to the risk. However, in Texas, Ohio, and Oklahoma, hundreds of earthquakes (magnitude even over M = 3 on the Richter scale) have been registered, and it has long been known that such earthquakes are a consequence of hydraulic fracturing (fracking) processes to stimulate oil production (Ellsworth 2013). From a content analysis of the media coverage about carbon capture and storage, Boyd and Paveglio (2014) concluded that framing in media articles not only brings the issue to the attention of citizens, but also that it can affect public views and opinions—an issue that is particularly relevant for controversial emerging technologies.

In the present study, we focus on coverage of gas drilling and induced seismicity in newspapers in The Netherlands, an issue that became highly controversial in society and the political arena in 2013 (Opperhuizen et al. 2019). News media coverage was limited for decades, but media attention expanded dramatically in 2013. In a previous analysis we showed that the sentiment of newspaper articles changed substantially at this tipping point (Opperhuizen et al. 2019). In an agenda-setting study, we showed that the change in media reporting interacted with a change in the political debates about gas drilling and gas drilling policy (Opperhuizen et al. 2019). In the present study, we aim to answer the question: *how do quantitative changes in the volume of media attention on emerging risks of earthquakes relate to changes in the content of media reporting at sentence level*? We study quantitative changes in journalists' use of particular subtopics at the detailed sentence level rather than headlines, paragraphs, or full articles. As we collected over 2000 media articles, the raw dataset at sentence level easily exceeded 120,000 entries. So, we created a big dataset for which human coding was realistically not feasible. As we wanted to study the subtle changes over time, both over the entire period of study and between specific years, we applied supervised machine learning (SML) to extract subtopics from the content of journalistic articles. Thus, we aim to show that this approach, as proposed by Margolin (2019), is fruitful in the communication field for analysing big data relating to observational content analysis without testing pre-formulated hypotheses based on human coding.

2 Analytical framework and approach

2.1 Framing and subtopics at sentence level

Matthes (2009), Matthes and Kohring (2009), and Cacciatore et al. (2016) all argued that framing has different theoretical understandings and has been conceptualized and

operationalized in various ways in the literature. According to Entman (1993a, b), frames transform information about an issue and tell the reader which elements are meaningful according to the information provider. In this paper, we follow the sociological school of Gamson and Modigliani (2017), where frames are the backbone of a storyline or a central idea that provides meaning. Furthermore, we adopt Lörcher and Neverla's (2015) terminology and use subtopic as the level below topic in newspaper articles. In the present study, the topic is the situation of gas drilling and its related induced seismicity. Individual subtopics do not entail all aspects of frames as defined by Entman (1993a, b, p. 52) *in such a way as to promote a particular problem definition, causal interpretation, moral evaluation, and/or treatment recommendation.* Subtopics emphasize mainly the *what* element in communications to audiences rather the *how* and the *why* of the issue. However, when subtopics systematically co-occur across a text in newspaper articles, they may be connected and be parts of a frame (Miller 1997; Matthes and Kohring 2009). In their work, Semetko and Valkenburg (2000) distinguished two kinds of frames utilized by the media to cover issues in the political arena. First, they identified episodic frames, referring to the description of the specific issue at hand. Second, in their article on media coverage of European politics, they identified five thematic frames utilized by TV and newspaper journalists throughout the more than 4000 stories that they studied. *Attribution to responsibility* was the most prominent generic frame that they identified, followed by *conflict*, *economic consequences*, *human interest*, and *morality.* In other studies, additional generic frames have sometimes been identified (Matthes and Kohring 2009) or another sequence of the prominent thematic frames (Dan and Raupp 2018).

Goffman (1974) argued that reframing can occur in media at any time when incongruent information becomes available and new meaningful elements arise about the situation or the issue. Previously, it was indicated that changes in media's issue attention do not usually follow a cyclical pattern (Lörcher and Neverla 2015), as described initially by Downs (1972). In dynamic issue-attention studies in news media, increases in the number of publications are usually used as a quantitative measure. Reframing can be interpreted as a dynamic qualitative change in media reporting, meaning that the introduction of new subtopics sometimes leads to others being replaced. Stanyer and Mihelj (2016) presented an overview of typology and periods of longitudinal studies in three prominent communication and media journals. They concluded that there are very few studies researching changes over time and urged media and communication scientists to combine the strengths of quantitative and qualitative research to provide a more sensitive understanding of dynamics in communication. Dynamics should be understood here as temporal changes, as well as continuity, in reporting the frequency of an issue and its subtopics in newspapers. Longitudinal news media studies usually entail many articles and are prone to changes in linguistics as well as to changes in the writers of the content. Human-based coding of hundreds or thousands of articles is costly and time-consuming, apart from many other methodological limitations such as the bias introduced by the human coders themselves (Van Gorp 2007). Therefore, Su et al. (2017) suggested analysing sentiment and topics from (social) media datasets by using hybrid methods that combine human and computer techniques.

2.2 Content analysis with supervised machine learning (SML)

Content analysis of media reporting is usually carried out by human-based coding of particular words judged to be suitable for the relevant study (Lewis et al. 1999). Content

analysis is nowadays often assisted by computer algorithms, enabling analysis of more significant numbers of newspaper articles through computer-based techniques (Riffe et al. 2019). The application of SML or other human–computer hybrid models creates new opportunities to analyse the ever-growing amount of publicly available content from digitalized newspapers or other media outlets, platforms, websites, and social media for which content analysis based on human coding has created significant challenges (Weare and Lin 2000). Although Walter and Ophir (2019) advocated the use of unsupervised machine learning methods using an inductive mixed-model computational approach. However, they also listed a set of limitations of unsupervised methods which make them less suitable for longitudinal studies and analysis of tipping points in communication. Su et al. (2017) reported that the strengths of human- and computer-based coding could be capitalized by applying supervised content analysis tools. SML is a relatively new technique that allows analysis of publicly available big data, which could hardly be analysed by applying traditional methods of content analysis (Weare and Lin 2000; Lewis et al. 1999). SML can bridge the gap between traditional thematic and automatic content analysis, according to Scharkow (2013). It is also a promising technique for longitudinal studies (Su et al. 2017). The hybrid, human–computer-based technique enables the computer to learn from a set of human-coded training documents (Zhang et al. 2010; Sebastiani 2013). With SML, a modified type of inductive coding can be applied in cases where information is fragmented or where no results of previous studies have been generated that provide codes for content analysis (Elo et al. 2013). Some limitations of hybrid, human–computer-based techniques are also mentioned, as rules 'learned' and linguistic patterns of a particular study cannot be transported directly to another big data analysis. For every new study, a new human-coded training set needs to be developed. SML requires sufficiently large datasets because subsets of the dataset are needed as the corpus to train the algorithms in the machine learning approach (Kim et al. 2017). In longitudinal studies, the training set should sufficiently represent the various episodes of the whole period, because, otherwise, linguistic and other changes in the data providers may not be represented in the training set, thereby creating biased results during the final analysis.

3 Materials and methods

3.1 Gas drilling case

We analyse the volume and the content of articles about gas drilling over 25 years in one local and four national newspapers. This case provided a compelling example to analyse the dynamics of issue attention in media and the changes in the content of reporting over a long period of time, because the risk was an emerging risk. The fact that it was an emerging risk provides the opportunity to follow its development longitudinally in both media reports and actual risk events. In addition, this case is especially interesting because the negative consequences (earthquakes) appear at regional level, whereas the whole country benefits from the gas drilling. In the case of a local risk, media framing is extra meaningful, given that only people living in the region surrounding the risk get first-hand information

about the risk, whereas others in the country depend on media reports; thus, media can play an important role in the dissemination of the risk, making it interesting to study.

Gas drilling has generated benefits of more than €280 billion for The Netherlands (Vlek 2018) since it started in the 1960s. Earthquakes were a new risk issue in this region of The Netherlands, a risk that increased gradually during the last decades, both in strength and in frequency, as described in detail by Vlek (2018). In the northern part of the country, citizens had already been experiencing very mild earthquakes (up to M = 3.5) for more than two decades. In August 2012, an M = 3.6 earthquake struck the region. Then, in January 2013, the Dutch supervisory authority, State Supervision of Mines (SSM), stated that the emerging risk of earthquakes should be considered to be a safety issue and that a further increase in magnitude and adverse consequences could not be excluded (State Supervision of Mines 2013). In the period that followed, no further increase in the magnitude of the earthquake risk was observed (Vlek 2018).

3.2 Data collection

The newspaper articles were selected from the digital database LexisNexis NL. The research query used: *Gaswinning OR gasboring AND Groningen AND NOT Waddenzee*. Articles from five newspapers were selected for analysis: four national newspapers, *NRC Handelsblad*, *de Volkskrant*, *de Telegraaf*, and *Algemeen Dagblad*, and one local newspaper, *Dagblad van het Noorden.* The comprehensive newspaper database LexisNexis had articles available from 1990 to 2016 for the national newspapers but only articles from 1999 to 2016 for *Dagblad van het Noorden* because, before 1999, there is no digital archive publicly available for this newspaper. A total of 4113 articles resulted from this selection. However, *Dagblad van het Noorden* has geographical variants—*North*, *South*, *East*, and *West* editions—leading to many duplicates. After the removal of duplicates, a final dataset of 2265 relevant news articles resulted for the analysis.

3.3 Qualitative content analysis

The unit of analysis is a sentence, because this provides more detailed information than the headlines or an entire article and more context than single words. In total, 120,033 sentences were included for analysis. From the 2265 news articles, a training set of 102 articles was selected, entailing 3786 sentences (3% of the total) that were used for human coding. The sentences were inductively coded by two researchers. This generated subtopics for the labelled sentences (see Table 1). We included a subtopic in this study if the subtopic was present in more than 5% of the sentences. We chose this cut-off point, because the reliability of the predicted subtopics below 5% of the dataset decreased substantially.

In order to check inter-coder reliability, 5% of the body content was selected (Emmert and Barker 1989). The reliability coefficient of Cohen's kappa was κ 0.68, which is substantial, according to Landis and Koch (1999), and represents good observer agreement, according to Altman (1991).

Table 1 Codebook

Subtopic	Description	Examples
Safety issue	The sentences mention that earthquakes are a safety issue for people in the region; safety has to be the first priority ('safety first'); house renovations are necessary to prevent collapse or physical injuries to humans or deaths; or safety measures must be taken/have been taken	'Groningen people are in danger in their own house, when do the investigations provide clarity?' 'But you cannot keep Groningen residents locked up in unsafe houses'
Decision making	The sentences refer to the number of policy decisions on gas production, or to a decision that should be made/has been made (by the Minister of Economic Affairs) to reduce or increase gas production	'The cabinet has decided to close the gas tap in Groningen again' 'At the beginning of this year, Kamp decided to reduce gas production at Loppersum by 80 percent'
Physical hazard	The sentences refer to the physical consequences of gas drilling, mentioning things such as land subsidence, an earthquake, progression in earthquake magnitude, or the direct link between cause (gas drilling) and effect (earthquakes)	'The sharp increase in earthquakes is caused by more natural gas being extracted from the soil due to the cold winter' 'Gas extraction changes the natural balance in the soil, increasing the pressure along cracks in the earth'
Material damage	The sentences focus on the physical damage in the region on houses, buildings, and heritage sites (like churches); on the number of damage claims; or on compensation after an earthquake or procedures for damage compensation	'Between 1997 and 2000, a total of 444 claims were awarded, with an amount of more than 800,000 euros [converted] being paid' 'There are new cracks in the walls'
Citizens' feelings	The sentences refer to citizens' feelings of anger, sadness, hopeless, fear, and worry; to people being so angry that they take to the streets to protest; to emotional consequences like depression, insomnia, and anxiety attacks; or to the decline or lack of trust and incomprehension of political choices	'He sees the anxiety on the faces of the inhabitants' 'They express their concern, anger, and fear at the many meetings about earthquakes in village houses and halls'
Research and advice	The sentences focus on the need for research, or on research about potential earthquakes and their consequences, or on research that has led to advice in favour of a decision (i.e. to reduce gas production)	'The State Supervision of Mines recommends 40 percent less gas to be pumped up' 'A recommendation from the State Supervision of Mines states that gas extraction in the Groningen field must therefore be curtailed immediately'
Disadvantaged position of the region	The sentences address the negative and/or unequal consequence for the region or inhabitants of this region (compared) with other regions who mainly gain from the gas extraction	'If it had happened in the Randstad, the world would have been too small' 'Our province is pinched like the face of an adolescent'
Benefits	The sentences focus on the gas revenues for The Netherlands, i.e. mentioning billions of euro earned, or on the economic loss that a decline in gas production would cost the Dutch State. They highlight the importance of gas production from an economic perspective	'The consequences for the National Budget, spread over the next 3 years, are a total of 2.3 billion euros due to the reduction in natural gas revenues' 'This is injected with natural gas revenues of 11 billion euros annually'

Table 1 (continued)

Subtopic	Description	Examples
International relations	The sentences refer to the import (mostly from Russia) and export of gas or refer to the position of the Dutch gas market in relation to other countries in Europe	'However, gas from Russia is sensitive and we are not becoming more independent' 'With an annual basis of 20 billion cubic metres of imported gas, to be supplemented with gas from the Groningen field'
Apologies	The sentences capture the idea that the involved institutions should apologize or had already apologized for the damage caused by earthquakes and the lack of interest in protecting the local civilians against earthquakes	'I am going to say that I am sorry that thousands of people have been confronted with the effects of the earthquakes in Groningen' 'The big word is out: the government says "sorry" to the people of Groningen about gas extraction'
Communication	The sentences refer to the lack of communication messages from the involved institutions to inform citizens or to communication between citizens and institutions in general	'That is why we are screwing up communication' 'Consultation with residents took much more time than expected'
Gas supply	The sentences focus the amount of gas needed for the Dutch gas market or about the amount of gas that is still available for drilling in the Netherlands	'Only in a harsh winter, if more gas is needed for warm feet, can gas up to 31 billion be extracted in Groningen' Criticism of the minister is that the focus is on the security of gas supply
Safety versus cost	The sentences refer to the fact that there is, on the one hand, an economic benefit and, on the other hand, the risk of earthquakes	'It concluded there that safety up to 2013 was subordinate to the revenue from the Groningen gas, although it was already clear since 1993 that gas extraction caused earthquakes' 'Money cannot play a role when it comes to safety'
Governance structure	The sentences refer to the (changed) governance structure and the relationship between several public/private institutions or the distribution of (new) responsibilities, interdependencies, and power relations between the involved institutions. Also, all the sentences refer to 'the gas building' (in Dutch: *het gasgebouw*), which is the name of the risk governance network	'The Dutch government becomes the full owner of Gasunie's gas pipelines' 'It's a "closed and closed system"…focused on consensus'

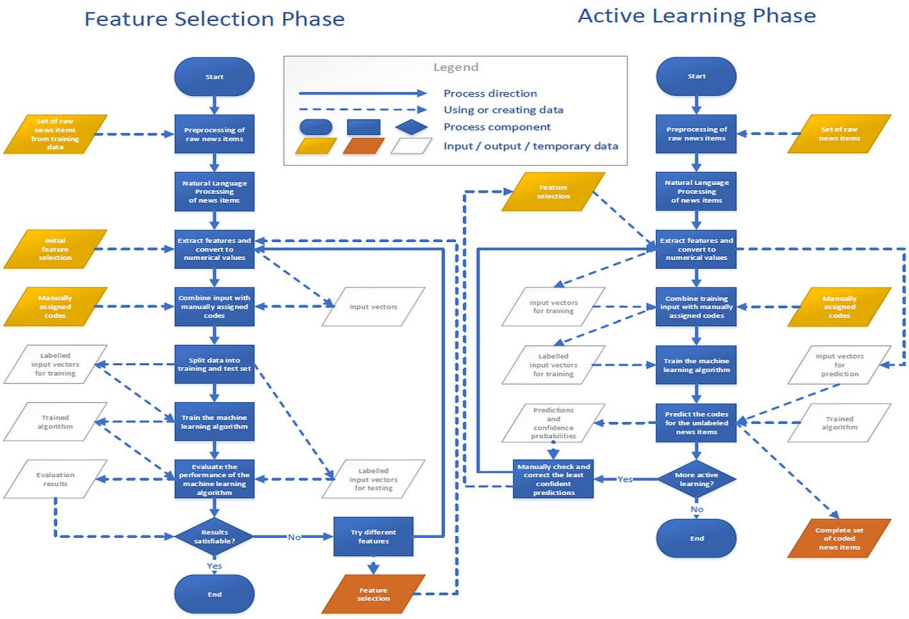

Fig. 1 The machine learning process

3.4 Supervised machine learning process

3.4.1 Train model, predict codes, and evaluate performance

The 3786 labelled sentences were exported from ATLAS.ti to XML documents, which formed the input for the SML (see Fig. 1 in "Appendix" for the process). Frog is used to tokenize the original sentences (Van den Bosch et al. 2007). The sequential minimal optimization (SMO) algorithm embedded in the LIBSVM tool is used to train the support vector machines (Chang and Lin 2011). The tokenized sentences are vectorized via the bag-of-words method (Zhang et al. 2010), which can be expanded with supervised machine procedures to include distributed word embedding (Rudkowsky et al. 2018). In this model, the order of the words in the sentence is no longer relevant; just the appearance of the words is essential. The bag-of-words model looks for the combination of all the words present in the bag, not just the presence of a single word. When possible, the bag-of-words model is complemented by other information derived from the text, such as sentiment information from dictionaries. However, SML algorithms use statistical methods and are not able to work with textual data as such. Hence, transforming the textual data into numerical data is a prerequisite. Processing steps are used to enable the conversion from textual data to numerical data that form the input for the SML algorithm.

3.4.2 Performance and reliability

According to Riffe et al. (2019), high reliability of a variable such as a subtopic indicates that it is manifest in the text, and such variables have higher reliability in comparison to full frames. In order to assess the performance of the employed algorithm, the labelled data (3786 sentences) were split into a 90% (3407 sentences) training set and a 10% (379 sentences) test set. Then, after training on the training data, the algorithm predicts the codes for the test data. Comparing the predicted codes against the manually assigned codes allows the computation of several performance measures. This process is repeated 10 times, as the split between the training set and the test set is random, and because of the relatively small number of labelled sentences. The performance measures are averaged to ascertain the accuracy of the employed method. Then, the standard deviation of these 10 performance measures is computed, as the spread between the 10 numbers gives a sense of how robust the results are. A low standard deviation means that performance is not dependent on having a lucky split between training and test data.

The performance measures used are computed for each of the codes, as each has a separately trained classifier. Whereas intercoder reliability is used for human-based coding, precision and recall can be used to indicate the performance of human–computer-based coding. Precision measures how many of the predictions are correct. Recall measures how many of the manually labelled codes are predicted. Precision and recall are often in a trade-off relation, meaning that one can improve at the expense of the other. It is easy to have very high precision by predicting only the code for a couple of easy sentences. It is even easier to have a very high recall, by merely predicting that the code is present in each sentence. To compare performances, it is therefore convenient to combine these two numbers; this third measure is called an F_1-score. It is the harmonic mean of precision and recall, and thus combines both characteristics into a single number (see Table 2). A low value for either precision or recall weighs down the F_1-score considerably.

Table 2 Subtopics' precision, recall, F_1-score, and SD. of F_1-score

Code of the news item	Average precision	Average recall	Average F_1-score	SD F_1-score
Safety issue	0.848	0.724	0.781	0.033
Decision making	0.863	0.680	0.761	0.060
Physical hazard	0.865	0.788	0.824	0.023
Material damage	0.857	0.689	0.764	0.039
Citizens' feelings	0.837	0.628	0.718	0.065
Research and advice	0.868	0.816	0.841	0.053
Disadvantaged position of the region	0.861	0.730	0.790	0.054
Benefits	0.856	0.721	0.783	0.044
International relations	0.848	0.756	0.799	0.056
Apologies	0.925	0.855	0.889	0.045
Communication	0.893	0.709	0.791	0.043
Gas supply	0.886	0.804	0.843	0.057
Safety versus cost	0.894	0.775	0.830	0.053
Governance structure	0.860	0.806	0.832	0.092

4 Results

We first discuss the results of the SML analysis, because the analysis itself is an important aspect of the study. Thereafter, the quantity of media messages in relation to the number of earthquakes is presented. Then, we present the content analysis, indicating the presence of the subtopics over time.

4.1 The media subtopics extracted with SML

In total, 120,033 sentences were included in the database to be analysed using SML. Each subtopic was generated independently of others, although more than one subtopic can originate from one sentence.

4.1.1 Accuracy and reliability

The performance measures used are computed for each subtopic generated from the training set. Each subtopic has a separately trained classifier (see 2). The analysis of newspaper coverage of gas drilling and earthquake risk from gas drilling generated 14 subtopics of content (see 1), based on codes that have a precision of over 0.8; this means that this prediction is more than 80% correct. The recall varies a bit more, as is usually the case for unbalanced datasets, ranging from 0.628 to 0.855. This means that the classifier has found between 62.8% and 85.5% of the sentences with a code.

The F_1-score then combines these numbers, and as this is the harmonic mean and not a normal average. The standard deviation in the next column is computed over the 10 F_1-scores for each code. In general, codes for data that are more unbalanced have a higher standard deviation. Also, codes that are more difficult for the algorithm to classify (e.g. citizens' feelings) have a higher standard deviation. However, the overall performance of the algorithm is robust, with standard deviations ranging from 0.023 to 0.065, and one outlier at 0.092 for *governance structure*.

Table 3 Frequencies and magnitudes of earthquakes related to number of articles published in five newspapers in The Netherlands

Period	1990–2002	2003–2008	2009–2012	2013–2015
No. of earthquakes	133	209	297	311
Mean of earthquakes	10	35	74	104
Max. strength	2.7	3.5	3.6	3.2
No. of newspaper articles	94	138	100	1933
National	39	42	34	711
Local	55	96	66	1222
Mean newspaper articles	7	23	25	644
Ratio newspaper articles/earthquakes	0.70	0.66	0.34	6.21
National	0.29	0.20	0.11	2.29
Local	0.41	0.46	0.22	3.93

4.2 Quantity of newspaper articles and the emerging risk

The total number of 2265 newspaper articles about gas drilling in the Netherlands did not show an equal volume distribution of attention over the 25 years of study, see Table 3. Nor did the distribution of the total number of newspaper articles synchronize with the frequency or the strength of the earthquakes.

In order to analyse the relationship between newspaper articles from local and national newspapers and the development of the risk over time in more detail, four periods are distinguished. Periods reflect either a substantive increase in the prominence of earthquakes (frequency or magnitude) or a change in the number of newspaper articles.

The first period ranges from 1990 to 2003, during which the annual frequency of earthquakes was relatively stable (a mean of 10/year) with no significant outliers. The actual hazard of every single earthquake in terms of strength was limited, with a magnitude within the range of 2.0–2.5 on the Richter scale. Although the risk events can be classified as relatively low, the permanence of earthquakes in the 1990s evolved into a new and chronic risk issue for the local community as a consequence of the gas drilling activity. As shown in Table 3, this new risk generated some media attention, with a mean of seven articles in a year, although this mean value underestimates actual attention because no numbers are available for the local newspaper before 1999. Despite this, the risk issue was only sporadically present in the newspapers. From 1999 onwards, the single local newspaper in the proximity of the gas drilling facilities had more coverage (55 articles in only 5 years from 1999 onwards) than the four national newspapers all together during the full period (39 articles all together in 14 years).

The second period is from 2003 to 2009, during which a substantive increase in earthquake frequency was registered, from a mean of 10 to a mean of 35 per year, and also with more earthquakes greater than M = 3. The most substantial earthquake had a magnitude of 3.5, 0.8 more than in the first period. The ratio between the annual number of articles and the number of registered earthquakes remained almost the same as in the first period, 0.7 in period 1 versus 0.65 in period 2. The increase in the volume of media attention is almost proportional to the increase in the earthquake risk. The local newspaper reported approximately twice as many articles as the four national newspapers all together (see Table 3). Interestingly, one of the national newspapers did not cover the gas drilling issue in this period at all.

In the third period, starting in 2009 and ending in 2012, a further increase in the frequency of the earthquake risk can be observed, from a mean of 35 to a mean of 74 earthquakes per year, as well as a small increase in the maximum magnitude up to M = 3.6 in 2012. Despite this, the mean number of news articles per year hardly exceeded that of the previous period, and the ratio between the annual number of articles and the number of registered earthquakes dropped to 0.34. This third period can be characterized as a period of relatively high risk but relatively low media attention.

The fourth and last period is from 2013 onwards. In this period, the frequency of earthquakes increased further to 104 per year, but the maximum magnitude of the earthquakes was lower than in the third period. However, during this period, both local and national newspapers reported frequently about the risks of gas drilling, and the mean number of articles per year became more than 20 times higher than in periods 2 and 3. The ratio between annual articles and annual earthquakes increased to 6.2 as a result, from 0.2 to 4 for the local newspaper (20 times more) and from 0.1 to 2.3 for national newspapers (23

Table 4 Subtopics registered in full dataset of sentences in newspapers during the four periods

	1990–2002	2003–2008	2009–2012	2013–2015
Safety issue	23	94	19	2170
Decision making	14	50	27	2044
Physical hazard	73	239	181	1496
Material damage	29	121	70	1600
Citizens' feelings	20	131	43	1299
Research and advice	14	79	16	687
Disadvantaged position of the region	9	73	9	596
Benefits	22	79	19	515
International relations	19	53	11	287
Apologies	9	37	8	279
Communication	7	40	5	266
Gas supply	8	39	5	240
Safety versus cost	5	36	6	165
Governance structure	4	42	4	119

times more). The increase in the reporting about the issue by local and national newspapers is not proportional to the increase in the actual risk of earthquakes.

4.3 Subtopics utilized in newspaper articles over time

All 14 subtopics are identified in the four periods of the study. However, the distribution of subtopics is different between these periods (see Table 4).

In the first period until 2003, the subtopic *physical hazard* was most prominent and was utilized in the newspapers approximately 10 times a year at sentence level (see Table 4). In the second period (2003–2008), the annual number of publications increased more, approximately 3.5 times. The use of all subtopics increased at least threefold. *Physical hazard* was the most prominent subtopic during this period also, although the increase in the use of subtopics was most prominent for *disadvantaged position of the region*, *communication*, and *governance structure*. *Safety*, *material damage*, *citizens' feelings*, and *benefits* were also subtopics in the second period, during which the actual risk of earthquakes increased substantially. In the third period, the annual number of articles was almost similar to that in the second period, and most subtopics were used less than in period 2. Only the use of *physical hazard* and *material damage* further increased in this period, during which the prominence of the actual risk also further increased. In period 4, after the Dutch SSM characterized gas drilling as a safety issue, *safety* and *decision making* became the most prominent subtopics used in the media, followed by *material damage* and *physical hazard*. Whereas the magnitude of the earthquake did not increase from period 3 to period 4, the use of *safety* increased 150 times and *decision making* 100 times. Also, the use of *disadvantaged position of the region* (88 times) and *communication* (71 times) increased more than the other items. Finally, although *decision making* became prominent in period 4, this is not observed for *governance structure*. Compared to period 2, for example, the annual use of *decision*

making increased by a factor of approximately 92, whereas, for *governance structure*, this increase was only 6.6.

In the limited number of articles published in the first three periods, *physical hazard* is the dominant subtopic. *Physical hazard* is still a vital subtopic in the fourth period (1496 hits), but slightly less than *material damage* (1600 hits). During the first three periods, the number of hits for *material damage* was only half that for *physical hazard.* Until 2013, the subtopic *benefits* paralleled the use of the subtopic *safety issue.* It may be related to *gas supply* and *international relationships*. *Safety versus costs* was extracted by SML as a separate subtopic. This subtopic, as well as *research and advice*, are also parallel in all four periods.

5 Discussion and conclusions

In this study, we aim to answer the question: *how do quantitative changes in the volume of media attention on emerging risks of earthquakes relate to changes in the content of media reporting at sentence level*?

We first discuss the subtopics relating to episodic framing, followed by the subtopics relating to thematic framing, to build up to an overall answer to the question.

5.1 Subtopics relating to episodic framing

During the first three periods, the increase in earthquake risk was more or less proportionally reflected in the Dutch newspapers by an increase in the volume of publications. The content did not change substantially during the two decades and showed similarities with media reporting in the US about fracking-induced seismicity (Fisk et al. 2017). In our study, we show that media utilized the subtopics *benefits*, *gas supply*, and *international relationships* to describe the economic perspective on gas drilling in The Netherlands, which can link to *physical hazard* and *material damage* as an adverse consequence. It may further link to the need for *research and advice* and *safety versus costs* evaluation. So, taken together and based on the description of the subtopics in Table 1, the media applied mainly an episodic frame to describe the economic activity in the northern part of The Netherlands. With regard to the definition of framing, according to Entman (1993a, b), newspapers describe (at least partially) the problem and address the causal interpretation but do not address moral evaluation and/or treatment recommendations.

5.2 Subtopics related to thematic frames

In the fourth period, the subtopic *safety issue* dominates. *Citizens' feelings* also became an important subtopic. *Safety*, as well as *citizens' feelings*, refer to thematic frames identified by Semetko and Valkenburg (2000), such as human interest and conflict. The use of these subtopics may coincide with that of the subtopic *apologies*. This links to Semetko and Valkenburg's thematic frame of morality. The use of *apologies* shows an almost similar pattern and frequency as *communication*. After *safety issue*, *decision making* became the second most prominent subtopic in the last period, a subtopic that relates to Semetko and Valkenburg's *responsibility*—the decision in the Dutch case being to reduce or increase gas production. The subtopic *decision making* also points to the thematic frame *conflict* and disagreement about the policy and politics regarding gas drilling risk. This entails the need

for alternative decisions and treatment. In contrast, the *risk governance structure* responsible for gas drilling-risk policy and politics received much less attention. Media in all four periods hardly use the latter subtopic, which shows the highest standard deviation in the SML analysis. Sentences addressing the morality of the subtopic *disadvantaged position of the local region* (Groningen) were initially covered mainly in the local newspaper but were also a prominent subtopic in the national media in the fourth period. Overall, during the fourth period, a disproportional increase in the usage of several subtopics, which can be linked to the thematic frames *conflict*, *morality*, *human interest*, and particularly *responsibility*, is observed. The thematic frame *economic consequences*, which dominated the first three periods, remained present as well as subtopics, and this links to episodic frames such as *physical hazard* and *material damage*. With the expansion of subtopics, a frame pattern can be constructed that is consistent with Entman's (1993a, b) definition of framing, as the combination of subtopics covers the particular problem definition of the earthquake risk in The Netherlands, the causality between gas drilling and earthquakes, the moral evaluation aspects of human interest and risk and benefits, as well as (the need for) treatment of the earthquake risk issue.

We conclude that the media content was reasonably stable, and media utilized mainly episodic frames as described by Semetko and Valkenburg (2000). However, for the disjointed quantitative increase in reporting (in 2013), the actual risk was not significant. The tipping point in media reporting did not follow an increase in the risk. The SSM's classification of a drilling-induced earthquake as a safety issue for society was the trigger. With the introduction of *safety* as an issue for society, drilling-induced seismicity became the responsibility of the government (Cvetkovich and Lofstedt 1999); this supports the outcome of our previous study about the interaction between media, policy, and politics of gas drilling (Opperhuizen et al. 2019) In the current study, we show that a governmental authority's classification of an issue as a safety problem for society can cause controversy in society, and this can be the source of the tipping point and the reframing of the content of media articles. It triggered media to introduce subtopics that can link to thematic frames in addition to episodic frames, as identified by Semetko and Valkenburg (2000). With the introduction of subtopics that link to the thematic frames, *responsibility*, *conflict*, *human interest*, and *morality*, journalists systematically and rapidly changed *what* was communicated to the audience about the issue of gas drilling and drilling-induced seismicity.

The emerging risk of earthquakes in The Netherlands was not covered in a timely fashion in the national media, thus leaving the general population almost unaware of the risk of gas drilling. An insufficient level of risk awareness among journalists may explain this, or it may indicate a lack of the minimum level of prominence required to achieve broad-scale coverage by media, as postulated by Neuman (1990). We conclude that conflict and controversy did not play a role in the media stories for an extended period, and the absence of controversy may help to explain the low (or under-) reporting. This finding supports and adds to Boyd and Paveglio's (2014) study concluding that framing in media articles not only brings the issue to the attention of citizens, but also can affect public views and opinions—a situation that is particularly relevant for controversial emerging technologies.

This study shows that, from 2013 onwards, media attention increased in a disjointed manner when content no longer related directly to the prominence of the risk only, but to the controversy about risk and benefits. We conclude that gas drilling was no longer a technical question and that a controversial value-loaded issue marked the tipping point of the reframing. The trigger for the tipping point and reframing most likely was the SSM report in January 2013. This report introduced a value conflict by mentioning gas drilling as a *safety issue*. We conclude that, when gas drilling was introduced as a value conflict, media

added emotionally loaded subtopics like *citizens' feelings, material damage*, and *disadvantaged position of the region.* This finding supports other studies finding that journalists create eye-catching messages relating to human interest and added news value (Carslaw 2008; Kitzinger 1999). The expansion of emotionally loaded subtopics coincided with the change in the sentiment of the articles, as we have reported previously (Opperhuizen et al. 2019).

In the process whereby more emotional items started to dominate the reporting, the beneficial aspects of gas drilling for society became relatively less important. *Decision making* is an appealing item for the media because it invites stakeholders' opinions and expressions of interest. The lack of *controversy/conflict* elements until 2013 may be a sufficient explanation of why national reporting was limited, whereas the introduction of the *safety* and *decision making* subtopics triggered the national media coverage. Consequently, intensified media attention on *safety* and *decision making* may stimulate the public discourse about the commitment, care, competence, and predictability of the government responsible for the safety of citizens (Neuman 1990).

Gas drilling, as studied here, provided a compelling case to analyse the dynamics of issue attention in the media and the changes in the content of reporting over time. It may serve as an example for other media attention patterns for other benefit-risk issues related to manmade technological activities such as hydraulic fracking for oil and gas.

6 Conclusion and discussion on supervised machine learning

In the current study, SML was successfully applied to a big dataset to analyse the content of media reporting about the risk and benefits of gas drilling over 25 years. By training the computer algorithm with a limited corpus of data, a large set of sentences could be analysed with the bag-of-words approach. Overall, we find support for Scharkow's (2013) statement that supervised text classification, which uses algorithms from machine learning, has the potential to become the standard method for the quantitative and the qualitative content analysis of big textual data.

We conclude that subtopics extracted with SML in a longitudinal study can be successful for *frame mapping* (Miller 1997) or to reconstruct *patterns* (Matthes and Kohring 2009), mainly when such maps or patterns can be meaningfully interpreted with frames described in the literature and with episodic information on the issue being studied.

6.1 Limitations

The first limitation of this study is that the SML focused on print media. Data from other traditional channels like radio and television and also digital media are not part of this study, although these channels have an essential role in issue attention on emerging risk. Future content analysis of digital media and the comparison between digital and traditional reporting in the case of emerging risk is therefore needed. The second limitation relates to the design of the study, which focuses on the emerging risk of earthquakes. Wardman and Löfstedt (2018), for instance, argued that risk dynamics are context dependent. Hence, the outcomes of this study are not automatically valid for other types of risks and require further study.

Appendix: The steps involved in supervised machine learning

First, the documents are cleaned so that they can be processed correctly in later steps. In particular, splitting the text into proper sentences can be challenging to do automatically, as a rule stating that sentences should end with a period or full stop does not always apply in news articles. For instance, headings are often not followed by a full stop, falsely giving the impression that the first sentence of the article directly follows the heading. Therefore, first, an extra empty line between the headline is inserted, and the first sentence is added to mark the difference. Another frequent formatting problem is the use of quotes, as the quote often ends after the period. This makes the full stop unnoticeable for the computer, and this also leads to the incorrect merging of two sentences.

The second step consists of running the documents through Frog (van den Bosch et al. 2007), a natural language pipeline for Dutch that extracts all kinds of linguistic information from the text. It gives the computer information about the text that would otherwise just be a string of characters. It splits the text into groups of characters that comprise words. These words group into sentences. After that, the words are categorized into word types (nouns, verbs) and labelled with their lemma (i.e. the dictionary form of the word). This step helps the algorithm to understand that different words can have the same meaning (be, are, is). Another difficulty of the Dutch language is that it is possible to form new words by combining two or more existing nouns. Humans easily recognize these compound nouns, but computers do not automatically recognize that a word is formed from two or more other words and hence see these compound nouns as entirely new words that are unrelated to their constituent nouns. This can be a problem because words with similar meanings will not be noticed as similar by the algorithm. That is why compound nouns (e.g. *aardbeving* (earthquake), consisting of *aard* (earth) and *beving* (quake), are labelled with a list of their constituent nouns.

The third step is feature selection; this involves determining what pieces of information the algorithm can use to predict the codes. The bag-of-words model is the starting point of feature selection. This means that, for each unique word in the dataset, its presence or absence in the current sentence is recorded. For example, if the dataset contains 5000 unique words, recording which words are present would result in a series of 5000 zeroes and ones, where only the words that are present are given a 1, and the rest will remain a 0. Besides the words, the presence or absence of word types (e.g. nouns, verbs), named entities (e.g. names of persons, organizations, or places), and compound noun constituents are also recorded. Furthermore, by using a Dutch sentiment lexicon, six additional pieces of information, or features, are encoded: the number of positive words, the number of negative words, the number of objective words, the number of subjective words, the total sentiment score of all the words in the sentence, and the total subjectivity score of all the words in the sentence. It should be noted that all presence or absence features are binary, whereas these last six features are not.

The three steps transform the textual data into a set of numerical vectors, one for each sentence. After that, the dataset is divided into a training set and a prediction set. The training set consists of the manually coded sentences, and the prediction set contains all the other sentences. The SML algorithm used for this task is a support vector machine (SVM) that has proved to be effective at text classification tasks (Suykens and Vandewalle 1999; Tong and Koller 1999). For each of the codes (Table 4), a separate SVM model is trained to predict, for each sentence in the prediction set, whether that particular code is present or not (see Fig. 1). As particular codes may be more difficult to predict than others, the SVM model not only yields the final prediction but also assigns a probability to the two scenarios (i.e. present and absent). The more certain the algorithm is, the higher the probability is of one of the two scenarios. If the algorithm does not have any clue, the probability of a particular code being present will be 50%, the same as the probability of that code being absent. For each code, these low probability cases are again manually annotated to provide the algorithm with more training data. This process can be seen as performing one round of active learning to improve prediction accuracy.

References

Altman, D.G.: Analysis of survival times. In: Altman, D.G. (ed.) Practical Statistics for Medical Research, p. 365. Chapman and Hall, London (1991)

Boyd, A.D., Paveglio, T.B.: Front page or "buried" beneath the fold? Media coverage of carbon capture and storage. Public Underst. Sci. **23**, 411–427 (2014)

Cacciatore, M.A., Scheufele, D.A., Iyengar, S.: The end of framing as we know it… and the future of media effects. Mass Commun. Soc. **19**, 7–23 (2016)

Carslaw, N.: Communicating risks linked to food: the media's role. Trends Food Sci. Technol. **19**, 14–17 (2008)

Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Trans Intell Syst Technol (TIST) **2**(3), 1–27 (2011)

Cvetkovich, T.T., Lofstedt, R.E.: Social trust: consolidation and future advances. In: Löfstedt, R., Cvetkovich, G. (eds.) Social trust and the management of risk, pp. 153–156. Earthscan Publications Ltd., London (1999)

Dan, V., Raupp, J.: A systematic review of frames in news reporting of health risks: characteristics, construct consistency vs. name diversity, and the relationship of frames to framing functions. Health Risk Soc. **20**, 203–226 (2018)

Downs, A.: Up and down with ecology-the issue attention cycle. Public Interest **28**, 38–50 (1972)

Ellsworth, W.L.: Injection-induced earthquakes. Science **341**, 1225942 (2013)

Elo, S., Kääriäinen, M., Isola, A., Kyngäs, H.: Developing and testing a middle-range theory of the well-being supportive physical environment of home-dwelling elderly. Sci. World J. **2013**, 945635 (2013). https://doi.org/10.1155/2013/945635

Emmert, P., Barker, L.L.: Measurement of Communication Behavior. Longman, New York (1989)

Entman, R.M.: Framing: toward clarification of a fractured paradigm. J. Commun. **43**, 51–58 (1993a)

Entman, R.M.: Projections of Power: Framing News, Public Opinion, and US Foreign Policy. University of Chicago Press, Chicago (1993b)

Fisk, J.M., Davis, C., Cole, B.: Who is at Fault? The media and the stories of induced seismicity. Policy Polit. **45**, 31–50 (2017)

Gruszczynski, M., Wagner, M.W.: Information flow in the 21st century: the dynamics of agenda-uptake. Mass Commun. Soc. **20**, 378–402 (2017)

Gamson, W.A., Modigliani, A.: The changing culture of affirmative action. In: Braungart, R.G., Braungart, M.M. (eds.) Research in Political Sociology, vol. 3, pp. 137–177. JAI Press, Greenwich, CT (2017)

Goffman, E.: Frame Analysis: An Essay on the Organization of Experience. Harvard University Press, Cambridge (1974)

Kahlor, L.A., Wang, W., Olson, H.C., Li, X., Markman, A.B.: Public perceptions and information seeking intensions related to seismicity in five Texas communities. Int. J. Disaster Risk Reduct. **37**, 101–147 (2019)
Kim, T., Cha, M., Kim, H., Lee, J. K., Kim, J.: Learning to discover cross-domain relations with generative adversarial networks. In: Proceedings of the 34th International Conference on Machine Learning, pp. 1857–1865 (2017)
Kitzinger, J.: Researching risk and the media. Health Risk Soc. **1**, 55–69 (1999)
Landis, J.R., Koch, G.G.: The measurement of observer agreement for categorical data. Biometrics **33**, 159–174 (1999)
Lewis, S.C., Zamith, R., Hermida, A.: Content analysis in an era of big data: a hybrid approach to computational and manual methods. J. Broadcast. Electron. Media **57**, 34–52 (1999)
Lörcher, I., Neverla, I.: The dynamics of issue attention in online communication on climate change. Media Commun. **3**, 17–33 (2015)
Margolin, D.B.: Computational contributions: a symbiotic approach to integrating big, observational data studies into the communication field. Commun. Methods Meas. **13**, 1–19 (2019)
Matthes, J.: What's in a frame? A content analysis of media framing studies in the world's leading communication journals, 1990–2005. Journal. Mass Commun. Q. **86**, 349–367 (2009)
Matthes, J., Kohring, M.: The content analysis of media frames: toward improving reliability and validity. J. Commun. **58**, 258–279 (2009)
Miller, M.M.: Frame mapping and analysis of news coverage of contentious issues. Soc. Sci. Comput Rev. **15**(4), 367–378 (1997)
Neuman, S.P.: Universal scaling of hydraulic conductivities and dispersivities in geologic media. Water Resour. Res. **26**, 1749–1758 (1990)
Opperhuizen, A.E., Schouten, K., Klijn, E.H.: Framing a conflict! How media report on earthquake risks caused by gas drilling: a longitudinal analysis using machine learning techniques of media reporting on gas drilling from 1990 to 2015. Journal. Stud. **20**, 714–734 (2019)
Riffe, D., Lacy, S., Watson, B.R., Fico, F.: Analyzing Media Messages, Using Quantitative Content Analysis in Research, 4th edn. Routledge, New York (2019)
Rudkowsky, E., Haselmayer, M., Wastian, M., Jenny, M., Emrich, Š., Sedlmair, M.: More than Bags of Words: Sentiment Analysis with Word Embeddings. Commun. Methods Meas. **12**, 140–157 (2018)
Schäfer, M.S.: Taking stock: a meta-analysis of studies on the media's coverage of science. Public Underst. Sci. **21**, 650–663 (2012)
Semetko, H.A., Valkenburg, P.M.: Framing European politics: a content analysis of press and television news. J. Commun. **50**, 93–109 (2000)
Scharkow, M.: Thematic content analysis using supervised machine learning: an empirical evaluation using German online news. Qual. Quant. **47**, 761–773 (2013)
Sebastiani, F.: Machine learning in automated text categorization. ACM Comput. Surv. (CSUR) **34**, 1–47 (2013)
Shoemaker, P.J., Schäfer, S.D.: Mediating the Message, pp. 781–795. Longman, White Plains (1996)
Stanyer, J., Mihelj, S.: Taking time seriously? Theorizing and researching change in communication and media studies. J. Commun. **66**, 266–279 (2016)
Su, L.Y.F., Cacciatore, M.A., Liang, X., Brossard, D., Scheufele, D.A., Xenos, M.A.: Analyzing public sentiments online: combining human-and computer-based content analysis. Inf. Commun. Soc. **20**, 406–427 (2017)
Suykens, J.A., Vandewalle, J.: Least squares support vector machine classifiers. Neural Process. Lett. **9**, 293–300 (1999)
State Supervision of Mines.: Aardbevingen in de provincie Groningen. Kenmerk **13010015** (2013)
Tong, S., Koller, D.: Support vector machine active learning with applications to text classification. J. Mach. Learn. Res. **2**, 45–66 (1999)
Van den Bosch, A., Busser, B., Canisius, S., Daelemans, W.: An efficient memory-based morphosyntactic tagger and parser for Dutch. LOT Occas. Ser. **7**, 191–206 (2007)
Van Gorp, B.: The constructionist approach to framing: bringing culture back in. J. Commun. **57**, 60–78 (2007)
Vlek, C.: Induced earthquakes from long-term gas extraction in Groningen, the Netherlands: statistical analysis and prognosis for acceptable-risk regulation. Risk Anal. **38**, 1455–1473 (2018)
Walter, D., Ophir, Y.: News frame analysis: an inductive mixed-method computational approach. Commun. Methods Meas. (2019). https://doi.org/10.1080/19312458.2019.1639145
Weare, C., Lin, W.Y.: Content analysis of the World Wide Web: opportunities and challenges. Soc. Sci. Comput. Rev. **18**, 272–292 (2000)

Wardman, J.K., Löfstedt, R.: Anticipating or accommodating to public concern? Risk amplification and the politics of precaution reexamined. Risk Anal. **38**, 102–1819 (2018)
Zhang, Y., Jin, R., Zhou, Z.H.: Understanding bag-of-words model: a statistical framework. Int. J. Mach. Learn. Cybernet. **1**, 43–52 (2010)